Telematika 2021

by Ira Rosianal Hikmah

Submission date: 19-Sep-2021 09:02PM (UTC+0700)
Submission ID: 1651884712
File name: Jurnal_Telematika_IND_Ira_Rosianal_Hikmah.docx (526.47K)
Word count: 4362
Character count: 27138

Klasifikasi *Data Mining* pada *Indian Liver Patient Dataset* (ILPD) Menggunakan Algoritma *Supervised Learning* dengan *Software* Orange

Ira Rosianal Hikmah

Program Studi Rekayasa Keamanan Siber, Politeknik Siber dan Sandi Negara
Jl. Raya Haji Usa, Putat Nutug, Ciseeng, Bogor, Jawa Barat, Indonesia
ira.rosianal@poltekssn.ac.id

***Abstract*— The development of data volume every day has resulted in the need for Data Mining to obtain valuable and useful data from the available data. Data Mining is often used in various fields, one of which is machine learning. One of the tools used for machine learning is supervised learning. In supervised learning, the data is divided into training and testing data with target attributes determined by the label or class. This study uses the Indian Liver Patient Dataset (ILPD) with 583 cases, which consists of 11 attributes, and the target attribute consists of two categories, namely patients with liver disorders (liver) and not (nonliver). The data consists of 583 cases divided into training and testing data by random sampling with a proportion of 80%: 20%. In the Data Mining process, sampling is done 100 times repetition. The supervised learning method in this study is the Decision Tree, Random Forest, SVM, Neural Network, Naïve Bayes, k-NN, and logistic regression. The Data Mining process uses Orange software with the test and score widget, making it easier and faster. This study also evaluates by generating a confusion matrix, the level of accuracy and precession, training and testing time, and the ROC curve.**

***Keywords*— Data Mining, supervised learning, training data, testing data, ILPD, Orange**

***Abstrak*— Perkembangan volume data setiap hari mengakibatkan perlunya *Data Mining* untuk mendapatkan data berharga dan berguna dari data yang tersedia. *Data Mining* sering digunakan dalam berbagai bidang, salah satunya adalah *machine learning*. Salah satu *tool* yang digunakan untuk *machine learning* adalah *supervised learning*. Pada *supervised learning*, data dibagi menjadi data *training* dan data *testing* dengan atribut target yang sudah ditentukan label atau kelasnya. Penelitian ini menggunakan *Indian Liver Patient Dataset* (ILPD) dengan 583 kasus terdiri atas 11 atribut, dan atribut targetnya terdiri atas dua kategori yaitu pasien ganguan hati (liver) dan tidak (nonliver). Data tersebut dibagi menjadi data *training* dan data *testing* secara *random sampling* dengan proporsi 80%:20%. Dalam proses *Data Mining*, dilakukan *sampling* 100 kali pengulangan. Metode *supervised learning* dalam penelitian ini adalah metode *Decision Tree, Random Forest, SVM, Neural Network, Naïve Bayes, k-NN,* dan *Regresi Logistik*. Proses *Data Mining* menggunakan *software* Orange dengan *widget test and score* sehingga memudahkan dan mempercepat proses *Data Mining*. Penelitian ini juga melakukan evaluasi dengan menghasilkan *confusion matrix*, tingkat akurasi dan presesi, waktu *training* dan *testing*, serta kurva ROC.**

***Kata Kunci*— *Data Mining*, *supervised learning*, data *training*, data *testing*, ILPD, Orange**

I. PENDAHULUAN

Data meningkat dari hari ke hari dan sangat sulit bagi seseorang untuk menganalisis volume data yang besar untuk pengambilan keputusan yang sempurna. Oleh karena itu, diperlukan *Data Mining* untuk mengekstrak data yang berharga dan berguna dari data yang tersedia [1]. *Data Mining* (penambangan data) merupakan proses penggalian informasi yang tersembunyi dari data yang ada dengan menemukan pola tertentu [2]. *Data Mining* sering digunakan di berbagai bidang ilmu, seperti teknologi basis data, pembelajaran mesin (*machine learning*), statistika, pengambilan informasi, jaringan saraf tiruan, kecerdasan buatan, dan sebagainya [3].

Terdapat empat *tools* yang digunakan untuk *Machine Learning*, yaitu *Supervised Learning, Reinforcement Learning, Active Learning*, atau *Unsupervised Learning*. Dua *tools* yang umum digunakan adalah *Supervised* dan *Unsupervised Learning* [4]. Terdapat tiga teknik *Data Mining* yaitu klasifikasi, regresi, dan *clustering*. Klasifikasi merupakan analisis data prediktif dimana data terbagi menjadi beberapa kelas yang sudah diketahui dan diberikan label. Regresi merupakan proses melihat hubungan dan pengaruh pada suatu *dataset*, yang terbagi menjadi satu variabel dependen dan satu atau lebih variabel independen. *Clustering* adalah analisis data yang mana data dibagi menjadi beberapa *cluster* berdasarkan karakteristik tertentu dan belum ada label dari *cluster* tersebut di awal analisis [5]. Klasifikasi dan regresi pada umumnya merupakan tipe pemodelan dari *supervised learning* sedangkan *clustering* biasanya merupakan tipe pemodelan *unsupervised learning*.

Ada beberapa strategi untuk mengetahui seberapa baik model yang terbentuk. Salah satunya adalah validasi silang. Metode validasi yang paling sederhana adalah metode *holdout* dimana data dibagi menjadi dua bagian yaitu data *train* dan data *testing*. Model dilatih dan dibentuk dengan data *training* dan dievaluasi pada data *testing*. Data *training* digunakan untuk menemukan parameter model sedangkan data *testing* dengan ukuran yang lebih kecil dibandingkan dengan data *traning*, digunakan untuk menguji keakuratan parameter atau model yang terbentuk [6].

Metode lain untuk menilai kinerja model adalah dengan analisis kurva ROC (*Receiver Operating Characteristic*). Berbeda dengan validasi silang, analisis kurva ROC hanya

berfungsi untuk kasus ketika variabel target memiliki tepat dua nilai yang berbeda (dua kategori) seperti benar dan salah atau positif dan negatif [7]. Pengukuran kebaikan model lainnya adalah dengan mengukur persentase prediksi dengan benar atau akurasinya [8]. *Confusion matrix* melakukan pengujian untuk memperkirakan objek yang benar dan salah [9]. Menurut Han dan Kamber, *confusion matrix* merupakan suatu alat yang memiliki fungsi untuk melakukan analisis apakah *classifier* tersebut sudah baik dalam mengenai *tuple* dari kelas yang berbeda. Nilai dari True Positive (TP) dan True Negative (TN) memberikan informasi ketika classifier dalam melakukan klasifikasi data bernilai benar sedangkan False Positive (FP) dan False Negative (FN) memberikan informasi ketika classifier salah dalam melakukan klasifikasi data [10].

Beberapa tahun yang lalu, terdapat banyak perangkat lunak *Data Mining* dikembangkan. Beberapa dari perangkat lunak tersebut tersedia secara bebas dan gratis [1]. Perangkat lunak *open source* merupakan perangkat lunak komputer dimana *source code* tersedia untuk umum. Pengguna dapat menggunakan, memeriksa, memperbaharui, dan mendistribusikan kepada siapa saja untuk digunakan. Salah satu perangkat lunak *Data Mining* yang *open source* adalah Orange yang dapat digunakan untuk visualisasi data dan analisis data [1]. Orange menyediakan pemodelan baik *supervised* maupun *unsupervised learning*. Bahkan dengan *widget test and score*, pengguna dapat menjalankan beberapa model sekaligus pada perangkat lunak tersebut. Orange juga menyediakan fitur evaluasi model seperti akurasi, presisi, waktu yang dibutuhkan untuk training dan testing, spesifisitas, dan sebagainya [11]–[14].

Penelitian ini menggunakan data *Indian Liver Patient* dari UCI-*Machine Learning Repository* dengan targetnya adalah menentukan diagnosa pasien apakah memiliki ganguan hati (liver) atau tidak (nonliver). Dengan data tersebut, dilakukan klasifikasi *Data Mining* dengan beberapa algoritma *supervised learning* dan menggunakan perangkat lunak *open source* Orange. Penelitian ini melakukan evaluasi pada model yang terbentuk dengan menghasilkan *confusion matrix*, tingkat akurasi, tingkat presisi, waktu *training*, dan waktu *testing*, serta menampilkan kurva ROC.

II. METODOLOGI

A. Data Penelitian

Penelitian ini menggunakan *Indian Liver Patient Dataset* (ILPD) dari repositori UCI – *Machine Learning Repository* [15]. Data tersebut terdiri atas 583 kasus, yang terdiri dari 416 pasien liver dan 167 pasien nonliver. Data tersebut dikumpulkan dari India. ILPD berisi 441 catatan pasien pria dan 142 pasien wanita. Setiap pasien yang usianya melebihi 89 tahun, terdaftar sebagai usia 90 tahun. Data ini terdiri atas 11 atribut (keterangan), yaitu:

1. Usia Pasien
2. Jenis Kelamin
3. TB (*Total Bilirubin*)
4. DB (*Direct Bilirubin*)
5. Alkfos (*Alkaline Phosphatase*)
6. SGPT (*Alamine Aminotransferase*)
7. SGOT (*Aspartate Aminotransferase*)
8. TP (*Total Proteins*)
9. ALB (*Albumin*)
10. Rasio A/G (*Albumin & Globulin Ratio*)
11. Kelas (1 menunjukkan pasien liver dan 2 menunjukkan pasien nonliver)

B. Model Supervised Learning dan Evaluasi

Terdapat empat *tools* yang digunakan untuk *Machine Learning*, yaitu *Supervised Learning*, *Reinforcement Learning*, *Active Learning*, atau *Unsupervised Learning*. Dua *tolos* yang umum digunakan adalah *Supervised* dan *Unsupervised Learning* [4]. Teknik *Data Mining* dapat dikelompokkan menjadi tiga tipe yaitu klasifikasi, regresi, dan clustering [5]. Penelitian ini menggunakan *tool Supervised Learning* dengan teknik klasifikasi dan regresi.

Klasifikasi *Data Mining* adalah penempatan objek-objek ke salah satu dari beberapa kelas yang telah ditentukan sebelumnya. Klasifikasi banyak digunakan untuk memprediksi kelas pada suatu label tertentu dengan membangun model berdasarkan data *training* dan menggunakan model tersebut untuk mengklasifikasikan data *testing* [16]. Pada penelitian ini terdapat 2 kelas, yaitu kelas 1 menunjukkan objek pasien liver dan 2 menunjukkan objek pasien nonliver. Regresi merupakan teknik *Data Mining* yang termasuk ke dalam *supervised learning* dan digunakan untuk memprediksi target numerik seperti regresi linier sederhana [17]. Namun jika variabel dependennya merupakan variabel kategori, maka *Regresi Logistik* dapat digunakan. Jika variabel dependennya hanya dua kategori maka disebut *Regresi Logistik* binomial namun jika lebih dari dua kategori maka disebut dengan *Regresi Logistik* multinomial [7].

Penelitian ini menggunakan. Terdapat 7 (tujuh) metode yang digunakan pada penelitian ini, yaitu metode *Decision Tree*, *Random Forest*, *Support Vector Machine (SVM)*, *Neural Network*, *Naïve Bayes*, *k-NN*, dan *Regresi Logistik*. Berikut ini merupakan penjelasan singkat untuk ketujuh metode yang digunakan pada penelitian ini:

1. *Decision Tree*

Decision Tree merupakan salah satu jenis metode *supervised learning*. Metode ini menyediakan sarana untuk mendapatkan model peramalan dalam bentuk aturan yang mudah diterapkan. Aturan-aturan ini memiliki bentuk IF-THEN, yang mudah diterapkan oleh pengguna. Pendekatan *Data Mining* ini menerapkan pendekatan sistem berbasis aturan [18]. Sebuah pohon terdiri dari *node* keputusan yang dihubungkan dengan cabang-cabang dari simpul akar sampai *node* daun (akhir). Setiap cabang akan diarahkan ke *node* lain atau ke *node* akhir untuk menghasilkan suatu keputusan [19].

2. *Random Forest*

Pada dasarnya merupakan *Decision Tree* berganda (*multiple Decision Tree*) [18]. Metode ini hampir sama dengan *Decision Tree classifier*, tetapi menambahkan beberapa keacakan pada model saat membuat pohon. Hal ini akan menghasilkan model yang lebih baik dan lebih stabil [1].

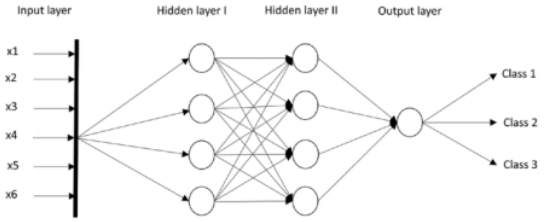
3. *Support Vector Machine* (SVM)

Metode ini pertama kali dikenalkan oleh Vladimir Vapnik dan mendapatkan popularitas karena banyak fitur menarik dan kinerja yang menjanjikan. SVM adalah algoritma *supervised learning* untuk klasifikasi dan regresi. Metode ini menggunakan bidang hiper untuk memisahkan titik data. Karena kinerja penyederhanaannya tinggi, metode ini dianggap sebagai *classifier* yang baik karena tidak membutuhkan pengetahuan sebelumnya [20]. SVM menghasilkan fungsi pemetaan *input-output* dari data *traning* yang telah diberi label. Fungsi tersebut dapat berupa fungsi klasifikasi atau regresi. Selain landasan matematika yang kuat dalam teori pembelajaran statistik, SVM menunjukkan kinerja yang sangat kompetitif di berbagai aplikasi seperti diagnosis medis, bioinformatika, pengenalan wajah, pemrosesan gambar, dan text mining. [18].

4. *Neural Network*

Neural Network cenderung bekerja lebih baik ketika ada hubungan yang rumit dalam data, seperti tingkat nonlinier yang tinggi. Dengan demikian metode ini cenderung menjadi model yang layak saat adanya ketidakpastian yang tinggi. Setiap node dihubungkan oleh busur ke node di lapisan berikutnya. Busur ini memiliki bobot, yang dikalikan dengan nilai node yang masuk dan dijumlahkan. Nilai simpul masukan ditentukan oleh nilai variabel dalam kumpulan data. Nilai node lapisan tengah adalah jumlah nilai node yang masuk dikalikan dengan bobot busur. Nilai simpul tengah ini pada gilirannya dikalikan dengan bobot busur keluar ke simpul penerus. *Neural Network* "belajar" melalui *loop* umpan balik. Untuk *input* yang diberikan, *output* untuk bobot awal dihitung. *Output* dibandingkan dengan nilai target, dan perbedaan antara *output* yang dicapai dan target diumpankan ke sistem untuk menyesuaikan bobot pada busur. Proses ini diulang sampai *network* dengan benar mengklasifikasikan proporsi data *training* yang telah ditentukan oleh pengguna (tingkat toleransi). Semakin baik kecocokan yang ditentukan, semakin lama waktu yang dibutuhkan *Neural Network* untuk berlatih, meskipun sebenarnya tidak ada cara untuk memprediksi secara akurat berapa lama waktu yang dibutuhkan model tertentu untuk belajar. [18]

Dalam bentuknya yang paling sederhana, *Neural Network* adalah fungsi linier. Dibutuhkan *input* berupa atribut dan mencari koefisien optimal (atau bobot) atribut untuk menghasilkan *output* yang sedekat mungkin dengan data eksperimen. Namun, *Neural Network* bisa jauh lebih rumit dan fleksibel bahkan bisa juga berbentuk nonlinier. Secara arsitektur, *Neural Network* memiliki banyak lapisan. Lapisan pertama adalah lapisan masukan (*input layer*), dan lapisan terakhir adalah lapisan keluaran (*output layer*). Di antara dua lapisan ini, bisa ada satu atau lebih lapisan tersembunyi (*hidden layer*). Informasi mengalir dari lapisan input ke lapisan tersembunyi, dan ke lapisan output [7]. Ilustrasi bentuk *Neural Network* sederhana dapat dilihat pada Gambar 1.



Gambar 1 Ilustrasi *Neural Network* sederhana [7]

5. *Naïve Bayes*

Naïve Bayes merupakan sekelompok algoritma probabilistik sederhana yang didasarkan pada Teorema Bayes (peluang bersyarat) [1]. Metode ini merupakan salah satu algoritma klasifikasi yang paling banyak digunakan [7].

6. *k-Nearest Neighbour* (k-NN)

Metode ini merupakan pengklasifikasi (*classifier*) sederhana yang menyimpan semua kasus yang tersedia dan kemudian menghasilkan kasus baru berdasarkan pengukuran serupa misalnya, fungsi jarak [1]. Orang cenderung mengambil keputusan atau mengambil tindakan berdasarkan saran dari orang-orang di sekitarnya. Itu terjadi dari waktu ke waktu dalam hidup kita bahwa kita menerima saran dari orang-orang di sekitar kita, terutama dari orang-orang yang sangat dekat dengan kita. Metode *Data Mining*, metode *k-Nearest Neighbour* (k-NN), merupakan cerminan dari momen kehidupan nyata tersebut. Di sini, k menunjukkan bahwa keputusan didasarkan pada k tetangga. Misalnya, di antara 10.000 sampel yang diketahui, ketika kasus baru A terjadi dan kita perlu memprediksi hasil A, kita menemukan 11 (biasanya ganjil) tetangga terdekat A dari sampel, dan berdasarkan hasil mayoritas dari 11 tetangga (k = 11 dalam kasus ini), hasil dari A diprediksi [7].

Ketika berbicara tentang *machine learning*, ada dua jenis pelajar (metode): pelajar yang bersemangat (*eager learner*) dan pelajar yang malas (*lazy learner*). Metode k-NN termasuk ke dalam *lazy learner* karena tidak membuat model terlebih dahulu, dan tidak menghasilkan serangkaian parameter atau aturan berdasarkan data *training*. Sebagai gantinya, ketika data penilaian masuk, *lazy learner* menggunakan seluruh rangkaian data pelatihan untuk mengklasifikasikan data penilaian secara dinamis. Dalam kasus k-NN, tetangga terdekat dari titik data penilaian dihitung secara dinamis menggunakan seluruh data *training*. Karena tidak ada set parameter yang dihitung sebelumnya, k-NN juga merupakan metode *Data Mining* nonparametric [7].

Pekerjaan utama dalam k-NN adalah menentukan tetangga yang terdekat. Diberikan titik data skor A, apa yang dapat didefinisikan sebagai tetangga terdekatnya? Dengan kata lain, bagaimana kita mengukur kedekatan? Salah satu cara untuk mengukur kedekatan adalah dengan menghitung jarak antara dua titik data. Jarak Euclidean, Manhattan dan Chebyshev adalah yang paling banyak digunakan untuk mengukur kedekatan dan untuk titik data numerik. Untuk atribut diskrit, jarak Hamming dapat digunakan [7].

7. *Regresi Logistik*

Regresi Logistik dapat dianggap sebagai kasus khusus dari regresi linier ketika yang diprediksi bersifat kategorik. *Regresi Logistik* adalah model statistik yang mampu memperkirakan probabilitas terjadinya suatu peristiwa. Beberapa data yang menarik dalam studi regresi mungkin berskala ordinal atau nominal. Karena analisis regresi memerlukan data numerik, maka dilakukan pengkodean pada variabel [18].

Misalkan 1 mewakili kasus ketika peristiwa terjadi dan 0 mewakili ketika peristiwa tidak terjadi. Oleh karena itu, $P(1)$ adalah peluang kejadian dan $P(0) = 1 - P(1)$. Misalkan terjadinya peristiwa tergantung pada variabel independen $X_1, X_2, \dots, X_n$. Dalam *Regresi Logistik*, log peluang adalah fungsi linier dari X_i dengan $i = 1, \dots, n$ yang digambarkan pada persamaan berikut: [7]

$$\ln\left(\frac{P(1)}{1-P(1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

Jika persamaan (1) diselesaikan, maka diperoleh:

$$P(1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \quad (2)$$

dengan β_i dengan $i = 0,1, \dots, n$ merupakan koefisien dari *Regresi Logistik* [18].

Setelah model terbentuk, dilakukan evaluasi. Ukuran yang digunakan dalam penelitian ini adalah tingkat akurasi dan presisi serta waktu *training* dan *testing*. Formula untuk tingkat akurasi dapat dihitung berdasarkan hasil *confusion matrix* sebagai berikut: [21]

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

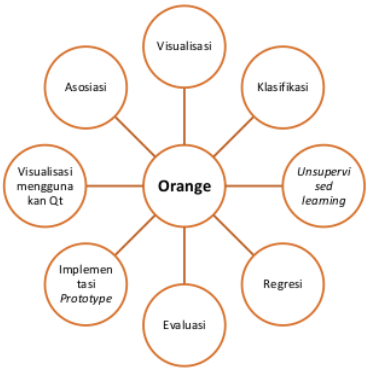
Selanjutnya, presisi merupakan nilai prediksi positif, atau dapat dihitung dengan formula berikut: [1]

$$Presisi = \frac{TP}{TP+FP} \quad (4)$$

dengan True Positive (TP) dan True Negative (TN) memberikan informasi ketika classifier dalam melakukan klasifikasi data bernilai benar sedangkan False Positive (FP) dan False Negative (FN) memberikan informasi ketika classifier salah dalam melakukan klasifikasi data [10]

C. *Software Orange*

Orange merupakan perangkat lunak *open source* untuk Machine Learning dan *Data Mining* yang ditulis dengan Bahasa Python. Orange dikembangkan oleh Laboratorium Bioinformatika, Fakultas Ilmu Komputer dan Informasi, Universitas Ljubljana [22]. Fitur yang disediakan pada *software* Orange dapat dilihat pada Gambar 2.



Gambar 2 Fitur *software* Orange [1]

Berikut ini beberapa *widget* pada *software* Orange yang digunakan dalam penelitian ini: [11]

1. *File*: klik dua kali pada *widget* ini untuk membuka dan memilih file data yang akan digunakan. *Output* dari *widget* ini adalah data.
2. *Select Columns*: klik dua kali untuk menentukan atribut target pada data. Pada *widget* ini, yang di-*input* adalah data.
3. *Data Table*: klik dua kali untuk melihat data dalam bentuk *spreadsheet*. Pada *widget* ini, yang di-*input* adalah data.
4. *Test and Score*: klik dua kali untuk melakukan pemodelan, prediksi, dan evaluasi model. Pada *widget* ini yang di-*input*-kan adalah data dan *learner* (model) dengan *output*-nya adalah prediksi dan hasil evaluasi.
5. *Data Sampler*: klik dua kali untuk membagi data menjadi data sampel dan *out-of-sample* atau yang disebut dengan *remaining data*. Akan muncul tampilan dan dapat dipilih metode pembagiannya apakah *fixed proporsion, fixed sample size*, dsb. Pada *widget* ini, yang di-*input* adalah data. Jika pada *widget test and score*, diterapkan pengulangan *random sampling*, maka *widget* ini tidak dibutuhkan.
6. *Tree, Random Forest, Neural Network, Naïve Bayes dan kNN*: *widget* ini dapat digunakan dengan dua bentuk, yaitu untuk menghasilkan model dan sebagai *learner*. Jika *output* adalah suatu model, maka yang di-*input* adalah data. Namun, jika *output* sebagai *learner* maka tidak perlu ada *input* pada *widget* ini.
7. *SVM*: *widget* ini memiliki tiga *output* yaitu model, sebagai *learner*, dan *support vector*.
8. *Logistic Regression*: *widget* ini memiliki tiga *output* yaitu model, sebagai *learner*, dan *coefficients*.
9. *Confusion Matrix*: klik dua kali pada *widget* ini akan menampilkan tabulasi silang atau dalam hal ini disebut dengan *confusion matrix* yang berisi jumlah data yang *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) untuk setiap model/*learner* yang dijalankan.
10. *ROC Analysis*: klik dua kali pada *widget* ini akan menampilkan kurva ROC untuk setiap *classifier*/*learner*

dengan sumbu X merupakan spesifisitas dan sumbu Y merupakan sensitivitas.

D. Langkah-Langkah Penelitian

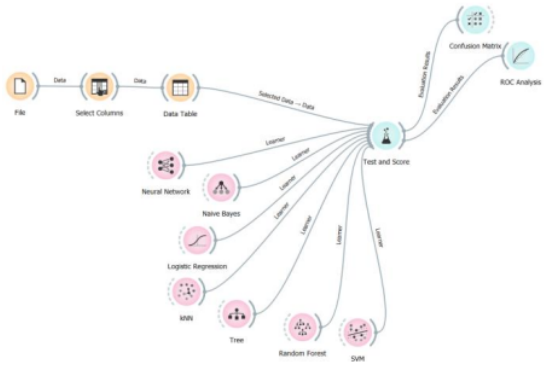
Secara umum, proses klasifikasi data mencakup dua langkah. Langkap pertama adalah membangun model atau aturan klasifikasi dengan mempelajari data training dengan label kelas terkait. Langkah kedua adalah menggunakan model tersebut untuk melakukan klasifikasi terhadap data testing untuk mendapatkan keakuratan dari model atau aturan klasifikasi tersebut [22].

Langkah-langkah secara rinci dalam penelitian ini merujuk pada beberapa langkah dalam penelitian Popy, sehingga didapatkan langkah penelitian sebagai berikut: [16]

1. Pengecekan dan pembersihan data
   Pada tahap ini, dilakukan pembersihan data untuk menghindari hal-hal seperti *incomplete*, *noisy,* dan *inconsistent*.
2. Proses integrasi data
   Pada tahap ini, dilakukan pengubahan atau konversi data ke jenis data yang diinginkan. Penelitian ini mengubah variabel target yaitu angka 1 dengan "Liver" dan angka 2 dengan "Non".
3. Menentukan target data
4. *Preprocessing*
   Tahap ini dilakukan sebelum proses *Data Mining*. Pada tahap ini dilakukan pengambilan sampel (*sampling*) secara acak untuk membaginya menjadi data *training* dan data *testing* dengan proporsi 80% : 20%. Penelitian ini melakukan pengulangan *sampling* sebanyak 100 kali.
5. Visualisasi
   Tahap ini bertujuan untuk melihat gambaran data secara grafis dan mengetahui ukuran seperti rataan, median, modus, kuartil, proporsi, dan sebagainya.
6. Proses *Data Mining*
   Tahap ini merupakan tahap pembentukan model supervised learning. Penelitian ini melibatkan 7 (tujuh) metode yaitu . Proses ini menggunakan *software* Orange dengan menggunakan fitur *test & score* [13]. Pada tahap ini juga dilakukan perhitungan evaluasi model. Ukuran evaluasi yang digunakan dalam penelitian ini adalah akurasi, presisi, waktu *training* dan waktu *testing*, serta menampilkan dan membahas kurva ROC-nya.
7. Pengetahuan
   Tahap ini merupakan tahap terakhir dalam penelitian dan diharapkan dapat memberikan pengetahuan baik dalam proses *Data Mining* dan pengetahuan mengenai penggunakan *software* Orang *Data Mining* agar dapat dimanfaatkan untuk kepentingan pemerintah dan masyarakat.

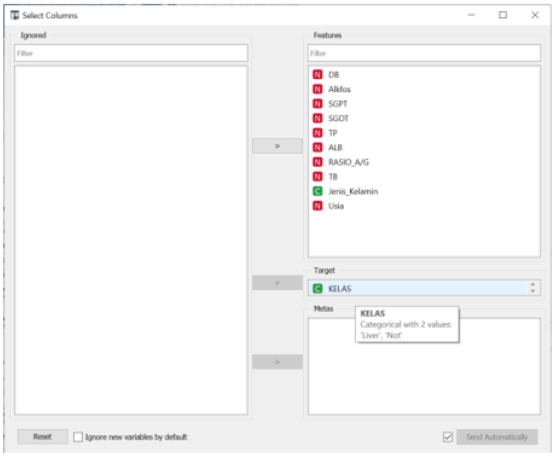
III. HASIL DAN PEMBAHASAN

Gambar 3 merupakan *workflow* penelitian dengan *software* Orange. Pada Gambar 3, terlihat bahwa terdapat tiga bagian, yang berwarna kuning, merah muda, dan hijau. Bagian yang berwarna kuning merupakan tahap persiapan data, yaitu *import* data, menentukan atribut bebas dan atribut target, dan membuat tampilan secara *spreadsheet*. Bagian yang berwarna merah menunjukkan metode *supervised learning* yang akan diterapkan. Penelitian ini menggunakan tujuh metode untuk menghasilkan *learner* yang kemudian pada bagian yang berwarna hijau akan diperoleh prediksi dan hasil evaluasi.



Gambar 3 *Workflow* penelitian dengan *software* Orange

Salah satu hal terpenting dalam data mining klasifikasi adalah menentukan atribut target. Ketika data di-*import*, maka semua variabel atau atribut akan dianggap sebagai atribut bebas. Oleh karena itu, *widget selected columns* perlu dilibatkan untuk memilih atribut target yang akan menjadi fokus penelitian. Proses penggunaan *widget selected columns* dapat dilihat pada Gambar 4.



Gambar 4 Melakukan pemetaan atribut bebas dan atribut target

Karena penelitian ini menggunakan *widget test and score* dan memilih pembagian data *training* dan data *testing* dengan proporsi 80%:20% secara *random sampling* dengan pengulangan sebanyak 100 kali, maka penelitian ini tidak menggunakan *widget Data Sampler* yang telah dijelaskan pada bagian sebelumnya. Tampilan pemilihan *random sampling* pada *widget test and score* dapat dilihat pada Gambar 5.

Gambar 5 Pemilihan *random sampling* dengan pengulangan 100 kali dan proporsi data *training* dan data *testing* adalah 80%:20%

Output dari *widget test and score* untuk hasil evaluasi dapat dilihat pada Gambar 6, 7, dan 8. Gambar 6 menunjukkan evaluasi untuk ketujuh model berdasarkan waktu yang dibutuhkan untuk *training* dan *testing*. Terlihat bahwa model k-NN membutuhkan waktu *training* paling cepat namun saat *testing* membutuhkan waktu yang paling lama. Waktu *training* paling cepat adalah model Naïve Bayes tetapi waktu *testing*-nya peringkat kedua dengan selisih 0,184 detik dibandingkan model *Decision Tree* yang waktu *testing*-nya adalah 0,016 detik. Model Regresi Logistik membutuhkan waktu yang cukup lama saat *training* yaitu sekitar 30,212 detik diikuti dengan model *Neural Network* yang membutuhkan waktu paling lama sekitar 101,701 detik. Walaupun *training* membutuhkan waktu yang lama, Regresi Logistik dan *Neural Network* membutuhkan waktu yang cukup cepat saat *testing*, masuk peringkat ketiga dan keempat. Model *Random Forest* dapat dikatakan membutuhkan waktu yang cukup cepat baik *training* maupun *testing*, dengan masing-masing waktunya adalah 2,893 detik (peringkat ketiga) dan 0,577 detik (peringkat kelima). Perlu dipahami bahwa hasil evaluasi berdasarkan waktu yang dibutuhkan untuk *training* dan *testing* model tergantung pada kemampuan dan spesifikasi *device*. Dalam penelitian ini, spesifikasi *device* dapat dilihat pada Tabel 1.

Evaluation Results

Model	Train time [s]	Test time [s]
Tree	9.388	0.016
Naive Bayes	1.039	0.190
Logistic Regression	30.212	0.216
Neural Network	101.701	0.500
Random Forest	2.893	0.577
SVM	4.665	0.605
kNN	0.796	1.015

Gambar 6 Hasil evaluasi berdasarkan waktu *training* dan *testing* yang dibutuhkan saat *Data Mining*

TABEL I
SPESIFIKASI *DEVICE* YANG DIGUNAKAN

Processor	Intel Core i7-10510U CPU @ 1.80GHz 2.30 GHz
RAM	16 GB
OS	Windows 10
Tipe Sistem	64-bit OS

Gambar 7 menunjukkan hasil evaluasi berdasarkan tingkat akurasi. Dapat dilihat bahwa model dengan tingkat akurasi tertinggi adalah model Regresi Logistik yang diikuti dengan *Neural Network, Random Forest, Naïve Bayes, k-NN, Decision Tree,* dan SVM.

Evaluation Results

Model	CA
Logistic Regression	0.718
Neural Network	0.711
Random Forest	0.702
Naive Bayes	0.669
kNN	0.668
Tree	0.663
SVM	0.656

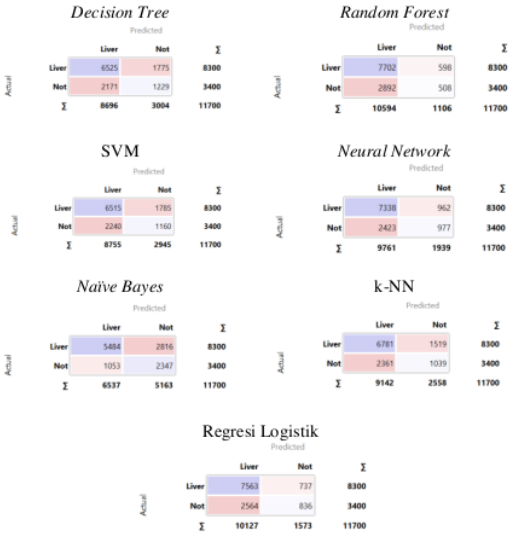
Gambar 7 Hasil evaluasi berdasarkan tingkat akurasi model

Gambar 8 menunjukkan hasil evaluasi berdasarkan tingkat presisi. Berbeda dengan tingkat akurasi yang mengukur kemampuan model memprediksi dengan benar, tingkat presisi disebut dengan nilai prediksi positif yaitu proporsi dengan hasil tes positif. Dapat dilihat bahwa model dengan tingkat presisi tertinggi adalah model *Naïve Bayes* yang diikuti dengan *Neural Network, Decision Tree,* Regresi Logistik, SVM, k-NN, dan *Random Forest*. Perhitungan tingkat akurasi maupun tingkat presisi berdasarkan pada *Confusion Matrix*. Hasil *Confusion Matrix* untuk ketujuh metode *supervised learning* pada penelitian ini dapat dilihat pada Gambar 9.

Evaluation Results

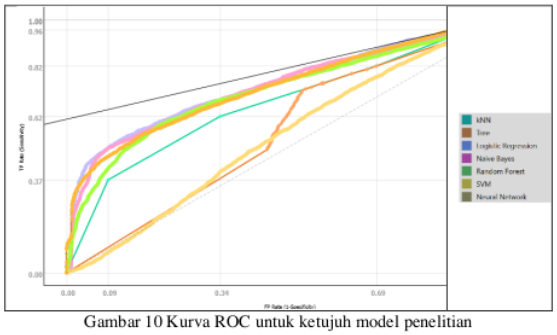
Model	Precision
Naive Bayes	0.839
Neural Network	0.752
Tree	0.750
Logistic Regression	0.747
SVM	0.744
kNN	0.742
Random Forest	0.727

Gambar 8 Hasil evaluasi berdasarkan tingkat presisi model



Gambar 9 *Confusion Matrix* setiap model penelitian

Penelitian ini juga melakukan analisis kebaikan model dengan analisis kurva ROC. Kurva ROC ketujuh model dapat dilihat pada Gambar 10. Terlihat bahwa empat model yang mendekati garis *performance* yaitu garis sebelah kiri yang berwarna hitam, yaitu model Regresi Logistik, *Neural Network, Random Forest*, dan *Naïve Bayes*. Kurva ROC mewakili hasil evaluasi berdasarkan tingkat akurasi model.



Gambar 10 Kurva ROC untuk ketujuh model penelitian

IV. SIMPULAN

Berdasarkan hasil penelitian, terlihat bahwa penggunaan *software* Orange untuk data *mining* dapat mempercepat waktu karena pengguna hanya tinggal membuat *workflow* dan melakukan konfigurasi dengan klik tanpa membuat sintaks *Python* walaupun *software* Orange memiliki *widget Python Script*. Pengguna dengan latar belakang non-IT khususnya pada data penelitian, di bidang kesehatan atau bidang lainnya seperti pertanian, dsb, dapat melakukan *Data Mining* Klasifikasi dengan *software* ini. Selain itu, terlihat bahwa waktu yang dibutuhkan untuk *training* dan *testing* model terbilang singkat. Selain itu, diperolah pula bahwa dalam penelitian ini dengan data yang digunakan (ILDP), empat model terbaik berdasarkan tingkat akurasi adalah Regresi Logistik, diikuti dengan *Neural Network, Random Forest*, dan *Naïve Bayes*.

DAFTAR REFERENSI

[1] R. Ratra dan P. Gulia, "Experimental Evaluation of Open Source Data Mining Tools (WEKA and Orange)," *Int. J. Eng. Trends Technol.*, vol. 68, no. 8, hal. 30–35, 2020.

[2] M. PhridviRaj dan C. GuruRao, "Data Mining - Past, Present and Future - A Typical Survey on Data Stream," *INTER-ENG ProcediaTechnology*, vol. 12, hal. 255–263, 2013.

[3] J. Han, M. Kamber, dan J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. USA: Morgan Kaufman Publisher, 2012.

[4] S. Shalev-Shwartz dan S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. New York: Cambridge University Press, 2014.

[5] M. Roos, "A data analysis demonstrator for managing customer experience in a partnering ventur," Stellenbosch University, 2019.

[6] T. Wendler dan S. Grottrup, *Data Mining with SPSS Modeler: TheORY, Exercises and Solution*, 2nd ed. Switzerland: Springer Nature Switzerland, 2021.

[7] H. Zhou, *Learn Data Mining Through Excel: A Step-by-Step Approach for Understanding Machine Learning Methods*. USA: Apress, 2020.

[8] A. Jose, M. Philip, L. T. Prasanna, dan M. Manjula, "Comparison of Probit and Logistic Regression Models in the Analysis of Dichotomous Outcomes," *Curr. Res. Biostat.*, vol. 10, no. 1, hal. 1–7, 2020, doi: 10.3844/amjbsp.2020.1.19.

[9] F. Gorunescu, *Data Mining Concepts, Model and Techniques Vol 12*. Berlin: Springer, 2011.

[10] J. Han dan M. Kamber, *Data Mining: Concepts and Techniques Tutorial*. San Francisco: Morgan Kaufman Publisher, 2001.

[11] B. Zupan dan Dems, "Introduction to data mining," 2018. doi: 10.1007/978-3-642-19721-5_1.

[12] Orange, "Orange Data Mining: Fruitful and Fun." [Daring]. Tersedia pada: https://orangedatamining.com/.

[13] J. Demsar dan B. Zupan, "Orange: Data mining Fruitful and Fun," *Informatica*, vol. 37, hal. 55–60, 2013.

[14] Orange Data Mining, "Orange Data Mining Library Documentation Release 3."

[15] UCI, "ILPD (Indian Liber Patient Dataset)." https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset) (diakses Agu 20, 2021).

[16] P. Meilina, "Penerapan Data Mining dengan Metode Klasifikasi Menggunakan Decision Tree dan Regresi," *J. Teknol. Univ. Muhammadiyah Jakarta*, vol. 7, no. 1, hal. 11–20, 2015.

[17] S. Sansgiry, M. Bhosle, dan K. Sail, "Factors That Affect Academic Performance Among Pharmacy Students," *Am. J. Pharm. Educ.*, vol. 70, no. 5, hal. Article 104, 2006.

[18] D. L. Olson dan D. Wu, *Predictive Data Mining Models*, 2nd ed. Singapore: Springer, 2020.

[19] Larose dan T. Daniel, *Discovering Knowledge in Data: An Introduction to Data Mining*. USA: John Wiley & Sons, 2005.

[20] S. Dash, S. K. Pani, S. Balamurugan, dan A. Abraham, *Biomedical Data Mining for Information Retrieval: Methodologies, Techniques and Applications*. USA: Scrivener Publishing, 2021.

[21] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation," Adelaide, 2007. [Daring]. Tersedia pada: http://arxiv.org/abs/2010.16061.

[22] A. Naik dan L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," *Procedia Comput. Sci.*, vol. 85, no. 2016, hal. 662–668, 2016, doi: 10.1016/j.procs.2016.05.251.

Ira Rosianal Hikmah, kelahiran Jakarta. Pendidikan Program Sarjana Program Studi Matematika FMIPA Universitas Indonesia dan melanjutkan Program Magister Program Studi Statistika Institut Pertanian Bogor. Minat riset terkait bidang matematika, statistika, aktuaria, manajemen risiko, dan *machine learning*.

Telematika 2021

ORIGINALITY REPORT

20%	19%	9%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	repository.unmuhjember.ac.id Internet Source	2%
2	www.neliti.com Internet Source	1%
3	repository.its.ac.id Internet Source	1%
4	Submitted to S.P. Jain Institute of Management and Research, Mumbai Student Paper	1%
5	yulianilive.wordpress.com Internet Source	1%
6	idec.ft.uns.ac.id Internet Source	1%
7	repository.bsi.ac.id Internet Source	1%
8	Submitted to Sriwijaya University Student Paper	1%
9	Amri Muliawan Nur, Muhammad Farid Wazdi, Bambang Harianto, Muhammad Farid Zaini.	<1%

	"Implementation of Naive Bayes Algorithm in Analyzing Acceptance of Poor Student Assistance", Journal of Physics: Conference Series, 2020 Publication	
10	actapress.com Internet Source	<1 %
11	text-id.123dok.com Internet Source	<1 %
12	www.iaeng.org Internet Source	<1 %
13	e-journal.unair.ac.id Internet Source	<1 %
14	Submitted to RDI Distance Learning Student Paper	<1 %
15	123dok.com Internet Source	<1 %
16	core.ac.uk Internet Source	<1 %
17	ejurnal.stmik-budidarma.ac.id Internet Source	<1 %
18	www.thenucleuspak.org.pk Internet Source	<1 %
19	"Biomedical Data Mining for Information Retrieval", Wiley, 2021	<1 %

	Publication	
20	Submitted to University College London Student Paper	<1 %
21	ejers.org Internet Source	<1 %
22	repository.uin-suska.ac.id Internet Source	<1 %
23	media.neliti.com Internet Source	<1 %
24	saucis.sakarya.edu.tr Internet Source	<1 %
25	www.kompas.com Internet Source	<1 %
26	Submitted to Universitas Jenderal Soedirman Student Paper	<1 %
27	repositorio.unesp.br Internet Source	<1 %
28	docs.mipro-proceedings.com Internet Source	<1 %
29	repository.uinjkt.ac.id Internet Source	<1 %
30	www.coursehero.com Internet Source	<1 %
31	bebasbanjir2025.wordpress.com	

	Internet Source	<1%
32	bmcbioinformatics.biomedcentral.com Internet Source	<1%
33	repository.nusamandiri.ac.id Internet Source	<1%
34	www.asnugroho.net Internet Source	<1%
35	www.detiksumsel.com Internet Source	<1%
36	www.idqidian.us Internet Source	<1%
37	Juan C. Alfaro, Juan A. Aledo, José A. Gámez. "Learning decision trees for the partial label ranking problem", International Journal of Intelligent Systems, 2020 Publication	<1%
38	Rozzi Kesuma Dinata, Hafizal Akbar, Novia Hasdyna. "Algoritma K-Nearest Neighbor dengan Euclidean Distance dan Manhattan Distance untuk Klasifikasi Transportasi Bus", ILKOM Jurnal Ilmiah, 2020 Publication	<1%
39	adoc.pub Internet Source	<1%

40	es.scribd.com Internet Source	<1 %
41	id.scribd.com Internet Source	<1 %
42	bigtha.blogspot.com Internet Source	<1 %
43	business.uc.edu Internet Source	<1 %
44	caridokumen.com Internet Source	<1 %
45	docplayer.info Internet Source	<1 %
46	id.123dok.com Internet Source	<1 %
47	ijcs.stmikindonesia.ac.id Internet Source	<1 %
48	isindexing.com Internet Source	<1 %
49	jemaatgpmsilo.org Internet Source	<1 %
50	jurnal.unsyiah.ac.id Internet Source	<1 %
51	link.springer.com Internet Source	<1 %

52	predatech.org Internet Source	<1%
53	repositori.unsil.ac.id Internet Source	<1%
54	repository.telkomuniversity.ac.id Internet Source	<1%
55	www.jatit.org Internet Source	<1%
56	ejournal.st3telkom.ac.id Internet Source	<1%
57	eprints.fri.uni-lj.si Internet Source	<1%
58	Jaejin Jang, Im.Y Jung, Jong Hyuk Park. "An effective handling of secure data stream in IoT", Applied Soft Computing, 2017 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		